УДК 631.3

А.Н. Серебровский
Институт проблем математических машин и систем НАН Украины, г. Киев
tsereb@voliacable.com

О технологии извлечения знаний из информационных ресурсов предметной области экспертной системы

Предлагается подход к созданию технологии извлечения знаний из текстовых электронных ресурсов, которые не имеют предварительного семантического описания. Технология предназначена для формирования и обновления базы знаний экспертной системы. Подход основан на онтологии и тезаурусе ограниченной предметной области. В технологии используется автоматизированная семантическая разметка текстовых документов. Указаны инструментальные средства технологических процедур. Данная технология позволяет снизить трудозатраты при создании и обновлении базы знаний экспертных систем.

Введение

Одной из необходимых компонент экспертных систем (ЭС) является подсистема извлечения знаний о предметной области (ПрО) из информационных ресурсов (ИР) с целью формирования и актуализации базы знаний ЭС. Здесь выделяются два этапа:

– предварительная обработка ИР, заключающаяся в отборе фрагментов текстов по проблематике ПрО, их сортировке, фильтрации, обобщении;

– формализация полученных знаний и их загрузка в базу знаний (БЗ).

Основным источником электронных документов в настоящее время является сеть Интернет. При этом значительная доля интернет-документов не обеспечивается семантическим описанием, что резко затрудняет реализацию целенаправленного извлечения знаний по конкретным аспектам ПрО. Трудоемкость предварительной обработки ИР, по различным данным, составляет до 80% всех затрат процесса извлечения знаний. Вследствие этого остаются актуальными разработки технологических приемов, позволяющих повысить эффективность процедур извлечения знаний из заранее не подготовленных информационных ресурсов (ИР). Одним из направлений подобных разработок является использование онтологий для семантического анализа естественноязыковых текстов [1-4]. Настоящая статья посвящена указанной проблеме.

Целью данной работы является разработка концепции автоматизированной технологии извлечения знаний из электронных текстовых ресурсов, позволяющей сократить трудозатраты на формирование и обновление БЗ ограниченной ПрО, не снижая при этом уровня полноты и достоверности извлекаемых знаний.

Данная технология должна быть основана на онтологии ПрО; использовать в качестве исходных ИР тексты электронных библиотек и Интернета, не имеющие семантического описания; включать в себя в качестве компонентов известные системные средства, которые поддерживают или полностью автоматизируют отдельные этапы извлечения знаний.

Постановка задач

Для достижении цели были поставлены задачи разработки и описания следующих технологических этапов:

– построения онтологии и тезауруса ПрО;
– семантической разметки электронных текстов, из которых будут извлекаться знания;
– извлечения знаний из размеченных текстов.

Описание технологии построения тезауруса и онтологии ПрО

Тезаурус и онтология ПрО строятся один раз при создании ЭС ПрО и затем многократно используются при ее эксплуатации. При необходимости дальнейшего расширения и уточнения тезауруса и онтологии применяется та же технология, что и при их построении. Изложение данной технологии согласуется (в основном) с концепцией А.С. Нариньяни [5]. Технология включает семь шагов, реализуемых экспертами и инженерами по знаниям при поддержке программных средств.

Шаг 1. Формирование комплекта текстов, покрывающих предметную область (КТПрО). Исходными материалами, из которых отбирается КТПрО, являются электронные ИР, в том числе тексты из специализированных журналов, справочников, отчетов, государственных и отраслевых стандартов, а также различные информационные материалы, выставленные в INTERNET. Отбор материалов может выполняться по: наименованиям журналов, статей; аннотациям; ключевым словам; классификационным признакам.

В качестве одного из оригинальных средств подготовки КТПрО может использоваться система поиска и анализа информации в Интернете «Галактика ZOOM» [6]. Данная система позволяет пользователю в диалоговом режиме создавать информационные портреты реальных объектов по текстовой информации, выполнять сравнительный анализ главных тем ИР и делать целевые выборки по заданному набору признаков.

Шаг 2. Составление словаря ПрО.

Эксперт просматривает содержание КТПрО, отмечая те лексические единицы, которые являются понятиями ПрО. Помеченные словоформы автоматически накапливаются, а затем упорядочиваются в алфавитном порядке, образуя словарь ПрО (СЛ).

Шаг 3. Формирование перечня терминов ПрО.

Эксперт фильтрует содержание СЛ, удаляя из него словоформы, связанные с жанровыми, стилистическими и другими особенностями данной ПрО. В результате формируется перечень слов и словосочетаний, являющихся терминами ПрО. Данный перечень обозначим «ТЕРМ».

Шаг 4. Формирование списка понятий ПрО.

Выполняются следующие действия:

– эксперт выполняет группировку терминов из перечня ТЕРМ. В каждую группу включаются термины, выражающие одно и то же понятие (синонимы);

– эксперт выбирает в каждой группе синонимов один термин, который будет представлять понятие этой группы в списке понятий онтологии (СП);

– автоматически «представитель» группы синонимов фиксируется в СП и ему присваивается уникальный ярлык (ТЭГ). Такой же ТЭГ получают соответствующие ему синонимы в перечне ТЕРМ.

В результате устанавливается соответствие между понятиями онтологии и их лексическими представлениями в тестовых документах ПрО.

По сути, совокупность СП и ТЕРМ представляют тезаурус ПрО.

Шаг 5. Классификация элементов СП в соответствии с базовыми семантическими категориями: объект, процесс, событие, свойство, значение и т.п.

В результате формируется СП «категорированный» (СПК).

Шаг 6. Установление базовых семантических связей между понятиями СПК.

Предварительно экспертами формируется базовый набор семантических отношений (часть – целое, частное – общее, объект – свойство, причина – следствие и т.п.). После этого между элементами СПК устанавливаются отношения из базового набора. Данная процедура является трудоемким процессом, требующая от экспертов значительных усилий. Остроту проблемы можно снизить, если придерживаться правила целесообразной достаточности, то есть ограничиться самыми существенными для функционирования ЭС связями между понятиями онтологии. В поддержку данного тезиса можно привести позицию А.С. Нариньяни: «Для большинства предметных областей моделью предметной области есть онтология с минимальной активной семантикой …» [5].

Процедура установления связей между понятиями в значительной степени определяется выбранным языком описания онтологий. К настоящему времени были разработаны и нашли применение различные языковые средства описания документов.

XML – (Extensible Markup Language) обеспечивает синтаксис для структурированных документов [7], [8].

XML Schema – добавляет к средствам XML возможности описания конкретных типов данных.

RDF – (Resource Description Framework) позволяет описать простую семантику произвольных ресурсов (понятия и отношения между ними), используя XML синтаксис [7].

RDF Schema – добавляет к средству RDF возможность описания иерархий понятий.

OWL – (Web Ontology Language) обеспечивает описание онтологий для Web ресурсов, а также для любых объектов. OWL разработан в трех модификациях (OWL LIFE; OWL DL; OWL FULL) [9]. OWL может рассматриваться в определенном смысле, как расширение RDF.

В качестве системного средства описания и редактирования онтологий может использоваться PROTEGE [10].

В результате формируется описание множества базовых отношений между конкретными понятиями ПрО.

Шаг 7. Добавление к полученным СПК и ОТБ понятий и отношений специфических для данной ПрО. Кроме того, в перечень ТЕРМ вносятся термины добавляемых понятий. Результатом является: расширенный список категорированных понятий (СПКр); расширенное описание отношений между понятиями (ОТБр); расширенный перечень терминов понятий (ТЕРМр).

Данные структуры образуют онтологию и тезаурус ПрО (рис. 1).

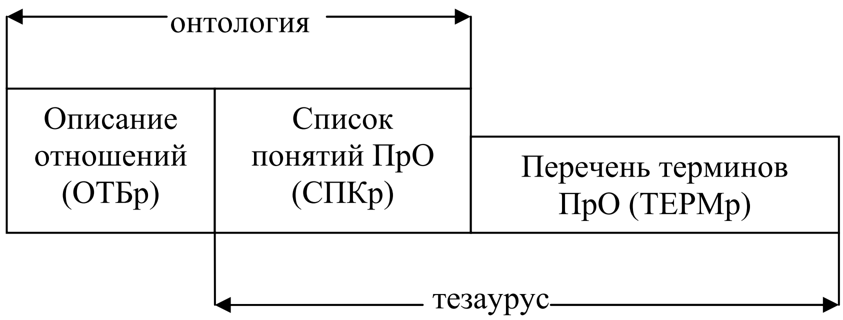


Рисунок 1 – Концепция комплекса «Онтология & Тезаурус»

Формально установленное соответствие между СПКр и ТЕРМр может быть представлено выражением

$$\forall c_i \setminus c_i \in СПК_p \quad \exists W_i = (w_{1_i}, w_{2_i}, ..., w_{k_i})..., \qquad (1)$$

где СПКр – расширенный список категорированных понятий онтологии ПрО;

W_i – класс синонимов из перечня ТЕРМ, соответствующих понятию c_i.

Семантическая разметка электронных текстов

Целью данного этапа является автоматизированное внесение в электронные тексты документов, составленных на естественном языке, формальных признаков отдельных понятий онтологии ПрО, характеризующих смысловое содержание документов. Семантическая разметка (СР) выполняется над ИР, которые пользователь отобрал как источники формирования и обновления БЗ ЭС. СР является подготовительным процессом для дальнейшего извлечения знаний и выполняется периодически по мере того, как возникает надобность актуализировать БЗ ЭС на основе новых ИР ПрО.

Можно выделить следующие шаги семантической разметки.

Шаг 1. Разбиение ИР на фрагменты.

Фрагментами ИР могут быть разделы документа, страницы и абзацы.

Цель разбиения документов – облегчение ориентировки пользователя в массиве текстовых фрагментов, которые будут получены в результате извлечения знаний.

Для фрагментирования может быть использован набор символов XML, вставляемых в текст для фиксации информации о его структуре [7], [8].

Примечание. Для малых по объему ИР фрагментирование выполнять не целесообразно.

Шаг 2. Первичная семантическая разметка ИР.

Данный этап выполняется программно, согласно следующему алгоритму.

Для каждого понятия $c_i (i = \overline{1, N})$ построенной онтологии из тезауруса выбирается соответствующий ему класс терминов синонимов $W_i = \{w_{j_i} \mid j = \overline{1, k_i}; i = \overline{1, N}\}$. Затем поочередно выполняется поиск этих терминов в размечаемом ИР. В случае, если в некотором фрагменте текста обнаружен хотя бы один термин $w_{m_i} \in W_i, m = (\overline{1, K_i})$, то данному фрагменту присваивается «ярлык» (ТЭГ), соответствующий понятию c_i, и поиск синонимов $w_{1_i}, w_{2_i}, ..., w_{k_i}$ продолжается в следующем фрагменте текста. После обработки всех фрагментов (поиска синонимов понятия c_i) процесс повторяется для очередного элемента онтологии (c_{i+1}).

В результате применения подобной процедуры ко всем понятиям онтологии, каждому j-му фрагменту размечаемого текста будет присвоено i_j ТЭГов, $i_j \in \overline{(0, N)}$, где N – количество понятий онтологии.

Шаг 3. Вторичная (дополнительная) разметка ИР.

На этом этапе выполняется дополнительная разметка, учитывающая онтологические отношения между понятиями.

Рассмотрим пример. Допустим, построена онтология ПрО «Оценка и анализ взрывоопасности на объектах типа бензоколонка». Допустим, что при построении онтологии были зарегистрированы понятия: «Перегрузка персонала» (ПП) и «Человеческий

фактор как источник опасности пожара» (ЧФ), при этом между ними было установлено и зафиксировано отношение:

$$<ПП\ подкласс\ ЧФ>, \qquad (2)$$

то есть перегруз персонала (во всех его формах) является подклассом причин опасности пожара, вызванных человеческим фактором.

Допустим, что некоторый фрагмент размечаемого текста ФрА содержит только лексические единицы, соответствующие понятию «ПП». Тогда при первичной разметке ему был присвоен ТЭГ «ПП». При вторичной разметке выполняется выявление всех отношений понятия «ПП», в том числе отношение (2). Исходя из логики этого отношения, ФрА содержит сведения о человеческом факторе и, следовательно, ему будет присвоен также ТЭГ «ЧФ». Это позволит при извлечении знаний о «Человеческом факторе» выявить фрагмент, содержащий сведения о «Перегрузке персонала».

Результатом семантической разметки является совокупность фрагментов, каждый из которых наряду с исходным текстом содержит набор ТЭГов, соответствующих понятиям онтологии, содержащимся во фрагменте. Формализованное описание семантической разметки имеет вид

$$«Размеченный\ текст» = \{mark\ \ frag_1\} \ldots, \qquad (3)$$

где $l = \overline{(1, L)}$, L – количество фрагментов текста;

$mark\ \ frag_l = [исходный\ фрагм._l \,\&\, \{ТЕГ_{i_l}\}]$;

$i_l = \overline{(1, I_l)}$, I_l – количество Тегов в $l_{ом}$ фрагменте.

Величина I_l может рассматриваться как характеристика информационной содержательности фрагмента текста.

Примечание. Первичная и вторичная разметки выполняются автоматически программными средствами. При этом, первичная разметка реализуется одним алгоритмом для всех понятий онтологии. Алгоритм вторичной разметки должен предусматривать столько логических ветвей, сколько типов отношений между понятиями онтологии должны быть учтены при разметке. Полученные тексты накапливаются в библиотеке размеченных текстов данной ПрО для последующего извлечения знаний по различным целевым запросам. Для этого могут использоваться репозитории, среди которых наиболее известными являются: UCI Knowledge Discovery in Databases Archive [11]; DEA Dataset Repository [12]; Frequent Item set Mining Dataset Repository [13]; XML Data Repository [8].

Извлечение знаний из размеченных текстов

Технологию извлечения знаний из различных текстов ПрО можно представить в виде следующих шагов:

1. Формирование запроса для целевого извлечения знаний.

Для формирования запроса используются не ключевые слова, а понятия онтологии ПрО. При этом целесообразно использовать язык описания запросов SPARQL [14].

2. Поиск по сформированному запросу в библиотеке размеченных текстов.

Выбор искомых фрагментов выполняется по критерию соответствия запроса пользователя и совокупности ТЭГов, описывающих понятийное содержание фрагментов.

3. Упорядочение найденных текстовых фрагментов.

Цель данного этапа – подготовить пакет найденных текстовых фрагментов к виду, удобному для последующей фильтрации. Упорядочение выполняется автоматичес-

ки по одному или нескольким ключевым признакам в зависимости от указания пользователя (инженера по знаниям).

Такими признаками могут быть: понятия онтологии с учетом их важности в запросе; информационная содержательность фрагмента I_l (3); дата происхождения информационного ресурса и др.

4. Фильтрация пакета найденных фрагментов.

Цель фильтрации – удаление повторов, малозначимых фрагментов, ошибочно найденных фрагментов (например, ошибок вызванных омонимией). Этап выполняется инженером по знаниям при сервисной программной поддержке.

5. Первичная формализация знаний, представленных в отфильтрованном пакете фрагментов.

Цель этапа – представить знания из каждого фрагмента в виде совокупности предложений на ограниченном естественном языке. По каждому фрагменту высвечиваются понятия онтологии в пределах одного абзаца текста. При этом учитываются категории понятий (объект, процесс, событие, свойство, значение и т.п.). Инженер по знаниям формирует предложение в соответствии с правилами ограниченного синтаксиса.

6. Описание внутреннего представления знаний формализованных предложений и загрузка в БЗ.

Описание данного этапа выходит за рамки данной статьи. Приведем лишь кратчайшее его содержание.

Инженер по знаниям последовательно в диалоговом режиме выводит предложения, сформированные в п. 5, и преобразует их в форму, принятую в модели знаний ПрО, после чего производится загрузка извлеченных элементов знаний о ПрО в БЗ. По каждому элементу выполняется автоматическая проверка повторяемости знания и его противоречивости с уже имеющимися знаниями. Результаты протоколируются и представляются инженеру по знаниям для дальнейшей интерпретации. Например, поступление одного и того же знания из разных независимых источников может повысить доверие к нему; изменение экстенсиональных знаний об одном и том же объекте в различные моменты времени может свидетельствовать о динамике ситуаций на объекте; несовпадение сведений об объекте в одном временном срезе ослабляет доверие к этим знаниям и требует дополнительной проверки и анализа. Особое внимание необходимо уделять изменениям интенсиональных знаний о ПрО, поскольку это свидетельствует либо о коренных изменениях в онтологии ПрО, либо о полном недоверии к одному из источников сведений.

Заключение

1. Предлагается концепция автоматизированной технологии извлечения знаний (АТИЗ) из информационных ресурсов (ИР), не имеющих предварительного семантического описания.

2. АТИЗ является одним из подходов снижения трудоемкости формирования базы знаний (БЗ) экспертной системы (ЭС), использующей ограниченную предметную область (ПрО).

3. АТИЗ основана на Онтологии и Тезаурусе ПрО, которые позволяют связывать метаданные ПрО с их лексическими представлениями, что является основой для автоматизированной разметки ИР с использованием метаданных онтологии. Последнее обстоятельство, в свою очередь, создает возможность в дальнейшем выполнять целенаправленный поиск знаний в ИР не по ключевым словам, а с использованием понятий ПрО.

4. Практическое значение АТИЗ в том, что она может быть использована как один из конкретных методических подходов при разработке подсистемы извлечения знаний из ИР, предназначенной для формирования и обновления БЗ ЭС.

Литература

1. Rogushina J. Ontological Approach to Domain Knowledge Representation for Informational Retrieval in Multiagent Systems / J. Rogushina, A. Gladun // International Jornal «Information Theories & Applications». – 2006. – Vol. 13, № 4. – P. 354-362.
2. Палагин А.В. К проектированию онтологоуправляемой информационной системы с обработкой естественно-языковых объектов / А.В. Палагин, Н.Г. Петренко // Математические машины и системы. – 2008. – № 2. – С. 14-23.
3. Невзорова О.А. Онтологическая поддержка методов решения задач семантико-синтаксического анализа текстов [Электронный ресурс] / О.А. Невзорова. – Режим доступа : http://www.raai.org/cai-08/files/cai-08_paper_ 234.doc
4. Гаврилова Т.А. Онтологический инжиниринг [Электронный ресурс] / Т.А. Гаврилова. – Режим доступа : http://www.big. spb.ru/publications/bigspb/km/ontolog_engeneering.shtml
5. Нариньяни А.С. ТЕОН-2: от Тезауруса к Онтологии и обратно / А.С. Нариньяни // Труды Международного семинара «Компьютерная лингвистика и интеллектуальные технологии». – М. : Наука, 2002. – Т. 1. – С. 199-154.
6. Бискулова А.С. «Галактика Zoom» – уникальная система поиска и аналитических исследований текстовой информации в Интернете / А.С. Бискулова, А.В. Антонов, П.В. Щедрый // Сборник трудов Восьмой Международной конференции «Интеллектуальный анализ информации». – К., 2008. – С. 93-102.
7. RDF/XML Syntax Specification (Revised). W3C Recommendation [Электронный ресурс]. – 2004. – Режим доступа : http://www.w3.org/TR/REC-rdf-syntax/
8. OWL Web Ontology Language Guide. W3C Recommendation [Электронный ресурс]. – 2004. – Режим доступа : http://www.w3.org/TR/owl-guide/
9. Horridge1 M.A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-DE Tools Edition 1.0/ Matthew Horridge1, Holger Knublauch2, Alan Rector1, Robert Stevens1, Chris Wroe11 [Электронный ресурс]. – 2004. – Режим доступа : http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf
10. UCI Knowledge Discovery in Databases Archive [Электронный ресурс]. – 2005. – Режим доступа : http://kdd. ics.uci.edu/
11. Dea Dataset Repository [Электронный ресурс]. – 2001. – Режим доступа : http://www.etm.pdx.edu/dea/dataset/Latest%20Changes%20v4.doc
12. Frequent Itemset Mining Dataset Repository [Электронный ресурс]. – 2000. – Режим доступа : http://fimi.cs.helsinki.fi/data/
13. XML Data Repository [Электронный ресурс]. – 2002. – Режим доступа : http://www.cs.washington.edu/research/xmldatasets/
14. SPARQL Query Language for RDF.W3C Recommendation [Электронный ресурс]. – 2008. – Режим доступа : http://www.w3.org/TR/rdf-sparql-query/

О.М. Серебровський

Про технологію виявлення знань з інформаційних ресурсів предметної області експертної системи

Пропонується підхід до створення технології виявлення знань з текстових електронних ресурсів, які не мають попереднього семантичного опису. Технологія призначена для формування і оновлення бази знань експертної системи. Підхід заснований на онтології і тезаурусі обмеженої предметної області. В технології використовується автоматизована семантична розмітка текстових документів. Вказані інструментальні засоби технологічних процедур. Дана технологія дозволяє зменшити трудовитрати при створенні і оновленні бази знань експертних систем.

A.N. Serebrovskiy

About Technology of Knowledge Extraction from the Informative Resources of Expert System Subject Domain

The approach to creation of knowledge extraction technology from text electronic resources which have not preliminary semantic description is offered. Technology is intended for forming and update of knowledge base of expert system. The approach is based on ontology and thesaurus of the limited subject domain. In this technology the automated semantic markup of text documents is used. The tools of technological procedures are indicated. The technology allows to decrease labour intensiveness at creation and update of knowledge base of expert system.

Статья поступила в редакцию 22.02.2010.